Reverse Engineering codul sursă al vaccinului BioNTech/Pfizer SARS-CoV-2

Bun venit! În acest post, vom arunca o privire caracter cu caracter la codul sursă al vaccinului BioNTech / Pfizer SARS-CoV-2 ARNm.

> *Update: după ce peste 1,7 milioane de oameni au vizitat această pagină, am decis să scriu o carte într-o temă similară. Pentru a deveni un cititor beta, vă rugăm să mergeți la această pagină pe Tehnologia vieții. Mulțumesc!*

> *Vreau să mulțumesc exprimate mare de oameni care au petrecut timp previzualizare acest articol pentru lizibilitate și corectitudine. Toate greșelile rămân ale mele, deși, dar mi-ar plăcea să aud despre ele rapid la bert@hubertnet.nl sau @bert_hu_bert*

Acum, aceste cuvinte pot fi oarecum zguduitoare - vaccinul este un lichid care se injectează în braț. Cum putem vorbi despre codul sursă?

Aceasta este o întrebare bună, așa că haideți să începem cu o mică parte din codul sursă al vaccinului BioNTech / Pfizer, cunoscut și sub numele de BNT162b2, cunoscut și sub numele de Tozinameran, cunoscut și sub numele de Comirnaty.

WHO
International Nonproprietary Names Programme

9/2020

Sequence / Séquence / Secuencia

GAGAAΨAAAC	ΨAGΨAΨΨCΨΨ	CΨGGΨCCCCA	CAGACΨCAGA	GAGAACCCGC 50
CACCAΨGΨΨC	GΨGΨΨCCΨGG	ΨGCΨGCΨGCC	ΨCΨGGΨGΨCC	AGCCAGΨGΨG 100
ΨGAACCΨGAC	CACCAGAACA	CAGCΨGCCΨC	CAGCCΨACAC	CAACAGCΨΨΨ 150
ACCAGAGGCG	ΨGΨACΨACCC	CGACAAGGΨG	ΨΨCAGAΨCCA	GCGΨGCΨGCA 200
CΨCΨACCCAG	GACCΨGΨΨCC	ΨGCCΨΨΨCΨΨ	CAGCAACGΨG	ACCΨGGΨΨCC 250
ACGCCAΨCCA	CGΨGΨCCGGC	ACCAAΨGGCA	CCAAGAGAΨΨ	CGACAACCCC 300
GΨGCΨGCCCΨ	ΨCAACGACGG	GGΨGΨACΨΨΨ	GCCAGCACCG	AGAAGΨCCAA 350
CAΨCAΨCAGA	GGCΨGGAΨCΨ	ΨCGGCACCAC	ACΨGGACAGC	AAGACCCAGA 400
GCCΨGCΨGAΨ	CGΨGAACAAC	GCCACCAACG	ΨGGΨCAΨCAA	AGΨGΨGCGAG 450
ΨΨCCAGΨΨCΨ	GCAACGACCC	CΨΨCCΨGGGC	GΨCΨACΨACC	ACAAGAACAA 500

Primele 500 de caractere ale ARNm BNT162b2. Sursa: Organizația Mondială a Sănătății

Vaccinul ARNm BNT162b2 are în centrul său acest cod digital. Are 4284 de caractere, deci s-ar potrivi într-o grămadă de tweet-uri. La începutul procesului de producție a vaccinului, cineva a încărcat acest cod la o imprimantă ADN (da), care apoi a convertit octeții de pe disc în molecule de ADN reale.

O mașină ADN Kilobaser Express

Dintr-o astfel de mașină vin cantități mici de ADN, care după o mulțime de prelucrare biologică și chimică ajung să fie ARN (mai multe despre care mai târziu) în flaconul de vaccin. O doză de 30 micrograme se dovedește a conține de fapt 30 micrograme de ARN. În plus, există un sistem inteligent de ambalare lipidică (grasă) care aduce ARNm în celulele noastre.

> *Update: Derek Lowe de celebrul blog-ul în curs de desfășurare de peste la Știință a scris un post cuprinzător "ARN Vaccinuri și lipidele lor", care explică îngrijit lipide și părți de livrare a vaccinurilor pe care eu nu sunt competente pentru a descrie. Din fericire Derek este!*
>
> *Update 2: Jonas Neubert și Cornelia Scheitz au scris această pagină minunat cu o mulțime de detalii cu privire la modul în care vaccinurile de fapt a lua produse și distribuite. Recomandat!*

ARN-ul este versiunea volatilă a "memoriei de lucru" a ADN-ului. ADN-ul este ca unitatea flash de stocare a biologiei. ADN-ul este foarte durabil, redundant intern și foarte fiabil. Dar, la fel ca computerele nu execută cod direct de pe o unitate flash, înainte de a se întâmpla ceva, codul este copiat într-un sistem mai rapid, mai versatil, dar mult mai fragil.

Pentru computere, acesta este RAM, pentru biologie este ARN. Asemănarea este izbitoare. Spre deosebire de memoria flash, memoria RAM se degradează foarte repede, cu excepția cazului în care are tendința de a avea grijă cu dragoste. Motivul pentru care vaccinul ARNm Pfizer/BioNTech trebuie depozitat în cele mai adânci congelatoare adânci este același: ARN-ul este o floare fragilă.

Fiecare caracter ARN cântărește de ordinul a $0{,}53 \cdot 10^{-21}$ grame, ceea ce înseamnă că există în jur de $6 \cdot 10^{16}$ caractere într-o singură doză de vaccin de 30 micrograme. Exprimată în octeți, aceasta este de aproximativ 14 petabytes, deși trebuie spus că aceasta constă în aproximativ 13.000 de miliarde de repetări ale acelorași 4284 de caractere. Conținutul informațional real al vaccinului este puțin peste un kilobyte. SARS-CoV-2 în sine cântărește în jur de 7,5 kiloocteți.

Update: În postarea originală aceste numere au fost dezactivate. Iată o foaie de calcul cu calculele corecte.

Cel mai scurt pic de fundal

ADN-ul este un cod digital. Spre deosebire de computere, care folosesc 0 și 1, viața folosește A, C, G și U/T ("nucleotidele", "nucleozidele" sau "bazele").

În computere stocăm 0 și 1 ca prezență sau absență a unei sarcini, sau ca curent, ca tranziție magnetică, sau ca tensiune, sau ca o modulare a unui semnal, sau ca o schimbare a reflexivității. Sau, pe scurt, 0 și 1 nu sunt un fel de concept abstract - ei trăiesc ca electroni și în multe alte întruchipări fizice.

În natură, A, C, G și U/T sunt molecule, stocate ca lanțuri în ADN (sau ARN).

În computere, grupăm 8 biți într-un octet, iar octetul este unitatea tipică de date prelucrate.

Natura grupează 3 nucleotide într-un codon, iar acest codon este unitatea tipică de procesare. Un codon conține 6 biți de informații (2 biți per caracter ADN, 3 caractere = 6 biți. Aceasta înseamnă $2^6 = 64$ de valori diferite ale codonului).

Destul de digital până acum. Când aveți îndoieli, mergeți la documentul OMS cu codul digital pentru a vă vedea singuri.

Unele lecturi suplimentare sunt disponibile aici - acest link ("Ce este viața") ar putea ajuta face sens de restul acestei pagini. Sau, dacă vă place video, am două ore pentru tine.

Deci, ce face acest cod DO?

Ideea unui vaccin este de a învăța sistemul nostru imunitar cum să lupte împotriva unui agent patogen, fără ca noi să ne îmbolnăvim efectiv. Din punct de vedere istoric, acest lucru a fost făcut prin injectarea unui virus slăbit sau incapabil (atenuat), plus un "adjuvant" pentru a speria sistemul nostru imunitar în acțiune. Aceasta a fost o tehnică analogă hotărâtă care implică miliarde de ouă (sau insecte). De asemenea, a fost nevoie de mult noroc și mult timp. Uneori a fost folosit și un virus diferit (fără legătură).

Un vaccin ARNm realizează același lucru ("educați sistemul nostru imunitar"), dar într-un mod asemănător cu laserul. Și vreau să spun asta în ambele sensuri - foarte îngust, dar și foarte puternic.

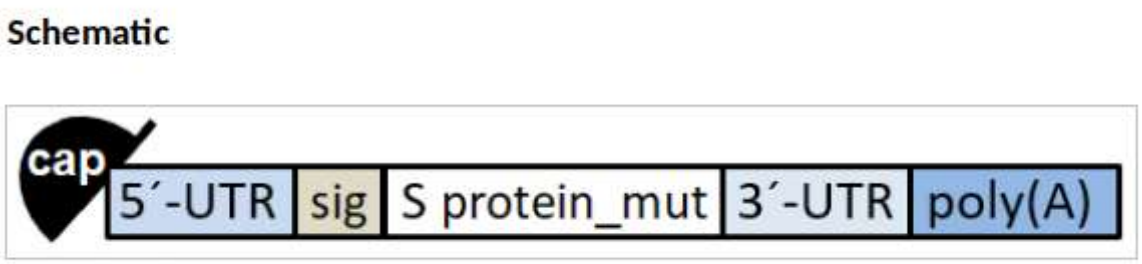
Deci, aici este modul în care funcționează. Injecția conține material genetic volatil care descrie celebra proteină SARS-CoV-2 "Spike". Prin mijloace chimice inteligente, vaccinul reușește să obțină acest material genetic în unele dintre celulele noastre.

Acestea încep apoi să producă proteine SARS-CoV-2 Spike în cantități suficient de mari încât sistemul nostru imunitar să intre în acțiune. Confruntat cu proteine Spike, și (important) spune-poveste semne că celulele au fost preluate, sistemul nostru imunitar dezvoltă un răspuns puternic împotriva mai multor aspecte ale proteinei Spike și procesul de producție.

Și asta ne duce la vaccinul eficient în 95%.

Codul sursă!

Să începem de la bun început, un loc foarte bun pentru a începe. Documentul OMS are această imagine utilă:



Acesta este un fel de cuprins. Vom începe cu "capacul", de fapt descris ca o pălărie mică.

La fel cum nu puteți doar să plonk opcodes într-un fișier de pe un computer și să-l executați, sistemul de operare biologic necesită anteturi, are linkers și lucruri cum ar fi convențiile de apelare.

Codul vaccinului începe cu următoarele două nucleotide:
`GA`

Acest lucru poate fi comparat foarte mult cu fiecare executabil DOS și Windows începând cu MZ, sau scripturi UNIX începând cu `#!`. Atât în viață, cât și în sistemele de operare, aceste două caractere nu sunt executate în niciun fel. Dar ei trebuie să fie acolo pentru că altfel nu se întâmplă nimic.

"Capacul" ARNm are o serie de funcții. Pentru unul, marchează codul ca provenind din nucleu. În cazul nostru, desigur, nu, codul nostru vine de la o vaccinare. Dar nu trebuie să spunem celulei asta. Capacul face ca codul nostru să pară legitim, ceea ce îl protejează de distrugere.

Primele două nucleotide sunt, de asemenea, ușor diferite din punct de vedere chimic de restul ARN-ului. În acest sens, are unele out-of-band de semnalizare pe ea.`GAGA`

"Regiunea netradusă cu cinci prime"

Unele lingo aici. Moleculele de ARN pot fi citite doar într-o singură direcție. Confuz, partea în care începe lectura se numește 5' sau "five-prime". Citirea se oprește la 3 'sau trei-prim sfârșitul.

Viața constă din proteine (sau lucruri făcute de proteine). Și aceste proteine sunt descrise în ARN. Când ARN-ul este transformat în proteine, aceasta se numește traducere.

Aici avem regiunea 5' netradusă ("UTR"), deci acest pic nu ajunge în proteină:
`GAAΨAAACΨAGΨAΨΨCΨΨCΨGGΨCCCCACAGACΨCAGAGAGAACCCGCCACC`

Aici întâlnim prima noastră surpriză. Caracterele ARN normale sunt A, C, G și U. U este, de asemenea, cunoscut sub numele de "T" în ADN. Dar aici găsim un Ψ, ce se întâmplă?

Acesta este unul dintre fragmentele extrem de inteligente despre vaccin. Corpul nostru rulează un sistem antivirus puternic ("cel original"). Din acest motiv, celulele sunt extrem de neentuziastice despre ARN-ul străin și încearcă din răsputeri să-l distrugă înainte de a face ceva.

Aceasta este oarecum o problemă pentru vaccinul nostru - trebuie să se strecoare dincolo de sistemul nostru imunitar. De-a lungul multor ani de experimentare, sa constatat că, dacă U în ARN este înlocuit cu o moleculă ușor modificată, sistemul nostru imunitar își pierde interesul. Pe bune.

Deci, în vaccinul BioNTech/Pfizer, fiecare U a fost înlocuit cu 1-metil-3'-pseudouridilil, notat cu Ψ. Partea cu adevărat inteligentă este că, deși acest înlocuitor Ψ placază (calmează) sistemul nostru imunitar, este acceptat ca un U normal de părțile relevante ale celulei.

În securitatea computerului știm, de asemenea, acest truc - uneori este posibil să transmitem o versiune ușor coruptă a unui mesaj care confundă firewall-urile și soluțiile de securitate, dar care este încă acceptată de serverele backend - care poate fi apoi hacked.

Acum culegem roadele cercetării științifice fundamentale efectuate în trecut. Descoperitorii acestei tehnici Ψ au trebuit să lupte pentru a-și finanța munca și apoi au fost acceptați. Ar trebui să fim cu toții foarte recunoscători și sunt sigur că premiile Nobel vor ajunge în timp util.

> *Mulți oameni au întrebat, ar putea virușii să folosească și tehnica Ψ pentru a ne bate sistemul imunitar? Pe scurt, acest lucru este extrem de puțin probabil. Viața pur și simplu nu are mașini pentru a construi 1-metil-3'-pseudouridyl nucleotide. Virușii se bazează pe mașinăria vieții pentru a se reproduce, iar această facilitate pur și simplu nu există. Vaccinurile ARNm se degradează rapid în corpul uman și nu există posibilitatea ca ARN-ul modificat de Ψ să se reproducă cu Ψ încă acolo. "Nu, într-adevăr, vaccinurile ARNm nu vor afecta ADN-ul "este, de asemenea, o lectură bună.*

Ok, înapoi la UTR 5 "." Ce fac aceste 52 de caractere? Ca totul în natură, aproape nimic nu are o funcție clară.

Când celulele noastre trebuie să *traducă* ARN-ul în proteine, acest lucru se face folosind o mașină numită ribozom. Ribozomul este ca o imprimantă 3D pentru proteine. Ingerează un fir de ARN și pe baza acestuia emite un șir de aminoacizi, care apoi se pliază într-o proteină.

Sursa: [Wikipedia utilizator Bensaccount]
(https://commons.wikimedia.org/wiki/File:Protein_translation.gif)

Aceasta este ceea ce vedem că se întâmplă mai sus. Panglica neagră din partea de jos este ARN. Panglica care apare în partea verde este proteina care se formează. Lucrurile care zboară în și în afară sunt aminoacizi plus adaptoare pentru a le face să se potrivească pe ARN.

Acest ribozom trebuie să stea fizic pe firul ARN pentru ca acesta să ajungă la locul de muncă. Odată așezat, poate începe să formeze proteine pe baza arnului suplimentar pe care îl ingerează. Din aceasta, vă puteți imagina că nu poate citi încă părțile în care aterizează mai întâi. Aceasta este doar una dintre funcțiile UTR: zona de aterizare ribozomă. UTR oferă "lead-in".

În plus, UTR conține și metadate: când ar trebui să se întâmple traducerea? Și cât de mult? Pentru vaccin, au luat cel mai "chiar acum" UTR pe care l-au putut găsi, luat din gena globinei alfa. Această genă este cunoscută pentru a produce robust o mulțime de proteine. În anii precedenți, oamenii de știință au găsit deja modalități de a optimiza și mai mult acest UTR (conform documentului OMS), deci acesta nu este chiar UTR-ul alfa globin. Este mai bine.

Peptida semnalului de glicoproteină S

După cum sa menționat, scopul vaccinului este de a obține celula pentru a produce cantități mari de proteină Spike a SARS-CoV-2. Până în acest moment, am întâlnit în mare parte metadate și chestii de "convenție de asteptare" în codul sursă al vaccinului. Dar acum intrăm pe teritoriul real al proteinelor virale.

Cu toate acestea, mai avem un strat de metadate de parcurs. Odată ce ribozomul (din splendida animație de mai sus) a făcut o proteină, acea proteină trebuie să meargă undeva. Acest lucru este codificat în "S glicoproteina semnal peptida (secvență lider extins)".

Modul de a vedea acest lucru este că la începutul proteinei există un fel de etichetă de adresă - codificată ca parte a proteinei în sine. În acest caz specific, peptida de semnal spune că această proteină ar trebui să iasă din celulă prin "reticulul endoplasmatic". Chiar și Star Trek lingo nu este la fel de fantezist ca acest lucru!

"Peptida semnalului" nu este foarte lungă, dar când ne uităm la cod, există diferențe între ARN-ul viral și cel vaccinal:

(Rețineți că, în scopuri de comparație, am înlocuit fantezie modificate Ψ cu un ARN U regulate)

```
            3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3
Virus:   AUG UUU GUU UUU CUU GUU UUA UUG CCA CUA GUC UCU AGU CAG UGU GUU
Vaccine: AUG UUC GUG UUC CUG GUG CUG CUG CCU CUG GUG UCC AGC CAG UGU GUG
                !   !   !   !   ! ! ! !     !   !   !   !   !           !
```

Deci, ce se întâmplă? Nu am enumerat întâmplător ARN-ul în grupuri de 3 litere. Trei caractere ARN alcătuiesc un codon. Și fiecare codon codifică pentru un anumit aminoacid. Peptida de semnal din vaccin constă *din exact* aceiași aminoacizi ca și în virusul în sine.

Deci, cum se face că ARN-ul este diferit?

Există $4^3 = 64$ codoni diferiți, deoarece există 4 caractere ARN și există trei dintre ele într-un codon. Cu toate acestea, există doar 20 de aminoacizi diferiți. Acest lucru înseamnă că mai multe codoni codifica pentru același aminoacid.

Life uses the following nearly universal table for mapping RNA codons to amino acids:

1st base	2nd base								3rd base
	U		C		A		G		
U	UUU	(Phe/F) Phenylalanine †	UCU	(Ser/S) Serine †	UAU	(Tyr/Y) Tyrosine †	UGU	(Cys/C) Cysteine †	U
	UUC		UCC		UAC		UGC		C
	UUA	(Leu/L) Leucine †	UCA		UAA	Stop (*Ochre*)[note 2]	UGA	Stop (*Opal*)[note 2]	A
	UUG		UCG		UAG	Stop (*Amber*)[note 2]	UGG	(Trp/W) Tryptophan †	G
C	CUU		CCU	(Pro/P) Proline †	CAU	(His/H) Histidine ‡	CGU	(Arg/R) Arginine ‡	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	(Gln/Q) Glutamine †	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	(Ile/I) Isoleucine †	ACU	(Thr/T) Threonine †	AAU	(Asn/N) Asparagine †	AGU	(Ser/S) Serine †	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	(Lys/K) Lysine ‡	AGA	(Arg/R) Arginine ‡	A
	AUG	(Met/M) Methionine †	ACG		AAG		AGG		G
G	GUU	(Val/V) Valine †	GCU	(Ala/A) Alanine †	GAU	(Asp/D) Aspartic acid ↓	GGU	(Gly/G) Glycine †	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	(Glu/E) Glutamic acid	GGA		A
	GUG		GCG		GAG		GGG		G

The RNA codon table (Wikipedia)

In this table, we can see that the modifications in the vaccine (UUU -> UUC) are all *synonymous*. The vaccine RNA code is different, but the same amino acids and the same protein come out.

If we look closely, we see that the majority of the changes happen in the third codon position, noted with a '3' above. And if we check the universal codon table, we see that this third position indeed often does not matter for which amino acid is produced.

So, the changes are synonymous, but then why are they there? Looking closely, we see that all changes *except one* lead to more C and Gs.

So why would you do that? As noted above, our immune system takes a very dim view of 'exogenous' RNA, RNA code coming from outside the cell. To evade detection, the 'U' in the RNA was already replaced by a Ψ.

However, it turns out that RNA with a higher amount of Gs and Cs is also converted more efficiently into proteins,

And this has been achieved in the vaccine RNA by replacing many characters with Gs and Cs wherever this was possible.

> *I'm slightly fascinated by the one change that did not lead to an additional C or G, the CCA -> CCU modification. If anyone knows the reason, please let me know! Note that I'm aware that some codons are more common than others in the human genome, but I also read that this does not influence translation speed a lot. UPDATE: A number of readers have pointed out that this change could prevent a "hairpin" in the RNA. You can try this out yourself on the RNAFold service.*
>
> *This marvelous article by Chelsea Voss goes into great depth on the RNA shape and contents of SARS-CoV-2.*

The actual Spike protein

The next 3777 characters of the vaccine RNA are similarly 'codon optimized' to add a lot of C's and G's. In the interest of space I won't list all the code here, but we are going to zoom in on one exceptionally special bit. This is the bit that makes it work, the part that will actually help us return to life as normal:

```
                  *   *
           L   D   K   V   E   A   E   V   Q   I   D   R   L   I   T   G
Virus:   CUU GAC AAA GUU GAG GCU GAA GUG CAA AUU GAU AGG UUG AUC ACA GGC
Vaccine: CUG GAC CCU CCU GAG GCC GAG GUG CAG AUC GAC AGA CUG AUC ACA GGC
           L   D   P   P   E   A   E   V   Q   I   D   R   L   I   T   G
            !     !!! !!       !   !       !   !   !   ! !
```

Here we see the usual synonymous RNA changes. For example, in the first codon we see that CUU is changed into CUG. This adds another 'G' to the vaccine, which we know helps enhance protein production. Both CUU and CUG encode for the amino acid 'L' or Leucine, so nothing changed in the protein.

When we compare the entire Spike protein in the vaccine, all changes are synonymous like this.. except for two, and this is what we see here.

The third and fourth codons above represent actual changes. The K and V amino acids there are both replaced by 'P' or Proline. For 'K' this required three changes ('!!!') and for 'V' it required only two ('!!').

It turns out that these two changes enhance the vaccine efficiency enormously.

So what is happening here? If you look at a real SARS-CoV-2 particle, you can see the Spike protein as, well, a bunch of spikes:

SARS virus particles (Wikipedia)

The spikes are mounted on the virus body ('the nucleocapsid protein'). But the thing is, our vaccine is only generating the spikes itself, and we're not mounting them on any kind of virus body.

It turns out that, unmodified, freestanding Spike proteins collapse into a different structure. If injected as a vaccine, this would indeed cause our bodies to develop immunity.. but only against the collapsed spike protein.

And the real SARS-CoV-2 shows up with the spiky Spike. The vaccine would not work very well in that case.

So what to do? In 2017 it was described how putting a double Proline substitution in just the right place would make the SARS-CoV-1 and MERS S proteins take up their 'pre-fusion' configuration, even without being part of the whole virus. This works because Proline is a very rigid amino acid. It acts as a kind of splint, stabilising the protein in the state we need to show to the immune system.

The people that discovered this should be walking around high-fiving themselves incessantly. Unbearable amounts of smugness should be emanating from them. And it would all be well deserved.

> *Update! I have been contacted by the McLellan lab, one of the groups behind the Proline discovery. They tell me the high-fiving is subdued because of the ongoing pandemic, but they are pleased to have contributed to the vaccines. They also stress the importance of many other groups, workers and volunteers.*

The end of the protein, next steps

If we scroll through the rest of the source code, we encounter some small modifications at the end of the Spike protein:

```
          V   L   K   G   V   K   L   H   Y   T   s
Virus:   GUG CUC AAA GGA GUC AAA UUA CAU UAC ACA UAA
Vaccine: GUG CUG AAG GGC GUG AAA CUG CAC UAC ACA UGA UGA
          V   L   K   G   V   K   L   H   Y   T   s   s
               !   !   !   !     ! !   !           !
```

At the end of a protein we find a 'stop' codon, denoted here by a lowercase 's'. This is a polite way of saying that the protein should end here. The original virus uses the UAA stop codon, the vaccine uses two UGA stop codons, perhaps just for good measure.

The 3' Untranslated Region

Much like the ribosome needed some lead-in at the 5' end, where we found the 'five prime untranslated region', at the end of a protein coding region we find a similar construct called the 3' UTR.

Many words could be written about the 3' UTR, but here I quote what the Wikipedia says: "The 3'-untranslated region plays a crucial role in gene expression by influencing the localization, stability,

export, and translation efficiency of an mRNA .. **despite our current understanding of 3'-UTRs, they are still relative mysteries**".

What we do know is that certain 3'-UTRs are very successful at promoting protein expression. According to the WHO document, the BioNTech/Pfizer vaccine 3'-UTR was picked from "the amino-terminal enhancer of split (AES) mRNA and the mitochondrial encoded 12S ribosomal RNA to confer RNA stability and high total protein expression". To which I say, well done.

The AAAAAAAAAAAAAAAAAAAAA end of it all

The very end of mRNA is polyadenylated. This is a fancy way of saying it ends on a lot of AAAAAAAAAAAAAAAAAA. Even mRNA has had enough of 2020 it appears.

mRNA can be reused many times, but as this happens, it also loses some of the A's at the end. Once the A's run out, the mRNA is no longer functional and gets discarded. In this way, the 'poly-A' tail is protection from degradation.

Studies have been done to find out what the optimal number of A's at the end is for mRNA vaccines. I read in the open literature that this peaked at 120 or so.

The BNT162b2 vaccine ends with:

```
                                        ****** ****
UAGCAAAAAA AAAAAAAAAA AAAAAAAAAA AAAAGCAUAU GACUAAAAAA AAAAAAAAAA
AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA AAAA
```

This is 30 A's, then a "10 nucleotide linker" (GCAUAUGACU), followed by another 70 A's.

There are various theories why this linker is there. Some people tell me it has to do with DNA plasmid stability, I have also received this from an actual expert:

"The 10-nucleotide linker within the poly(A) tail makes it easier to stitch together the synthetic DNA fragments that become the template for transcribing the mRNA. It also reduces slipping by T7 RNA polymerase so that the transcribed mRNA is more uniform in length".

The article "Segmented poly(A) tails significantly reduce recombination of plasmid DNA without affecting mRNA translation efficiency or half-life" also has a compelling description of how a linked can benefit efficacy.

Summarising

With this, we now know the exact mRNA contents of the BNT162b2 vaccine, and for most parts we understand why they are there:

- The CAP to make sure the RNA looks like regular mRNA
- A known successful and optimized 5' untranslated region (UTR)
- A codon optimized signal peptide to send the Spike protein to the right place (amino acids copied 100% from the original virus)
- A codon optimized version of the original spike, with two 'Proline' substitutions to make sure the protein appears in the right form
- A known successful and optimized 3' untranslated region
- A poly-A tail with a 'linker' in there

The codon optimization adds a lot of G and C to the mRNA. Meanwhile, using Ψ (1-methyl-3'-pseudouridylyl) instead of U helps evade our immune system, so the mRNA stays around long enough so we can actually help train the immune system.

Further reading/viewing

If you like this work, you can hire me to write about your scientific/technical/medical product as well!

In 2017 I held a two hour presentation on DNA, which you can view here. Like this page it is aimed at computer people.

In addition, I've been maintaining a page on 'DNA for programmers' since 2001.

You might also enjoy this introduction to our amazing immune system.

Finally, this listing of my blog posts has quite some DNA, SARS-CoV-2 and COVID related material.

As an update, the other up and coming vaccines are described in The Genetic Code and Proteins of the Other Covid-19 Vaccines

As a further update, there is now also a post describing the CureVac mRNA vaccine. The CureVac vaccine consists of mRNA that has not been modified, but instead has taken a leaf out of other parts of biology in hopes of making things work, and the post touches on those.

> *Update: after over 1.7 million people visited this page, I've decided to write a book in a similar theme. To become a beta reader, please head to this page on The Technology of Life. Thanks!*